コンピュータシステム用高速キャッシュメモリ技術

著者	元吉 真
号	51
学位授与番号	3715
URL	http://hdl.handle.net/10097/37383

	もとよし　まこと
氏　　名	元吉　真
授与学位	博士　(工学)
学位授与年月日	平成１８年９月１３日
学位授与の根拠法規	学位規則第４条第１項
研究科，専攻の名称	東北大学大学院工学研究科（博士課程） バイオロボティクス専攻
学位論文題目	コンピュータシステム用高速キャッシュメモリ技術に関する研究
指導教員	東北大学教授　小柳光正
論文審査委員	主査　東北大学教授　小柳光正　　東北大学教授　羽根一博 東北大学教授　寒川誠二　　東北大学教授　小林広明 東北大学助教授　田中徹

論文内容要旨

近年の半導体技術の進歩によって、コンピュータシステムの高性能化が急速に進んでいる。現在主流のプログラム内蔵式アーキテクチャーでは、処理スピードはデータパスと制御機能を持つCPU(central processing unit)の性能と、プログラム及びデータをメインメモリのRAM(random access memory)から取り出して収納する時間で決まる。このため、メモリシステムでは比較的小さい容量であるが高速動作可能なキャッシュをメインメモリのDRAM(dynamic random access memory)とCPUの間にバッファとして持つ階層構造が使われている。1990年代初めには、システムの高性能化及びRISC(reduced instruction set computer)型のアーキテクチャーの台頭により、高速で大容量のキャッシュが必要となり、Sun Microsystems、IBM、HP、DEC、SGI等に代表されるサーバーやエンジニアリングワークステーションメーカーでは、CPUチップ内蔵の１次キャッシュ、外付け２次キャッシュSRAM(static random access memory)、メインメモリ、ハードディスクという４層のメモリ階層構造を採り始めた。

CPUのクロック周波数が上がると、それに対応して、外付けの２次キャッシュの高速化が要求される。DRAMやSRAMは通常2.5～3年で次世代のプロセスに移行して４倍の容量、高速化が実現されるが、CPUの性能に合わせるため２次キャッシュは１年ごとに性能を上げることが要求されてきた。

本研究ではプロセス世代間のギャップを埋めるため、電流駆動能力の高いバイポーラトランジスタを付加したBICMOS SRAMや高駆動能力MOSトランジスタを入れた２次キャッシュSRAMデバイス及びプロセスの検討を行った。MOSトランジスタの電流駆動能力とSRAMの待機時消費電流に係るオフ電流はトレードオフの関係にある。待機時消費電流はLSIチップ内のゲート長分布下限近くのトランジスタのリーク電流で決まるため、LSIチップ内、ウェーハ内のゲート長バラツキを抑えることは重要

であり、バラツキに関して統計的に処理し、最適なゲート長(統計的な中央値)を用いたデバイス設計手法を考案するとともに、SRAMメモリセルレイアウトを幾何学的に分析して、ゲート寸法のバラツキの少ない最適な高速メモリセルレイアウトを導いた。

キャッシュSRAMは大規模なWebサーバーに使われるため、初期不良はもちろん、ソフトエラーにも高い信頼性基準が適用される。ソフトエラーは、半導体パッケージ中に不純物として含まれる放射性元素から出るα線が原因で、シミュレーションによる解析技術やソフトエラーレートの評価技術やデバイス構造の改善、使用材料の純度の向上によって、対策されたかに見えた。しかし、0.25μm世代になり素子の微細化に伴って電源電圧が下がり、記憶ノード容量が小さくなると、従来のソフトエラーのモデルからの予測値より2桁多いエラーが観測されたため、原因の究明を行った。0.25μmと0.18μmの8MbitSRAMを用い通常の海抜0mのフィールド試験に加えて海抜2000mでの高地試験、地下試験、熱中性子のシールド試験を行い、この結果から宇宙線が起源の熱中性子及び高速中性子によるものがそれぞれ全体の約3/4、1/4を占め、今まで問題にされていた放射性不純物元素起因は1%以下であることを明らかにした。また、この研究から、高速の中性子の入射によって6トランジスタSRAMのメモリセルがラッチアップを起こす可能性があることを明らかにし、デバイス構造の改善により、不良率が低減できることを世界に先駆けて確認した。また、従来の入力端子からのノイズによるラッチアップと異なり、LSIチップ内部の局所的な領域で電圧降下が起こり、ハードエラーまでには至らないことを明らかにした。

コンピュータシステムの高性能化の一方で、コンピュータ1台あたりの消費電力も増加している。データセンタは、Webサービス事業や大企業のコンピュータ・ネットワークの運用を行っているが、数百台から数千台規模のサーバーや関連機器が設置されており、空調の電力も含めてエネルギーコストの高騰が深刻な問題になっている。米国カリフォルニア州では電力が自由化されているが、情報化時代のコンピュータ、ネットワーク、通信システムが以前は予備電力だった分を消費しており、余剰電力が殆ど無い状態である。1998年12月8日のサンフランシスコ停電、2000年6月14日のベイエリアの停電では、特に社会的に影響が大きいデータセンタへの電力供給が優先的に行われたことから、情報機器の消費するエネルギーの問題が大きく取り上げられれるようになった。データセンタではコンピュータのメインメモリやネットワークデータの蓄積のために膨大な数のDRAMが使われており、2008年にはコンピュータシステムの約半分の電力をDRAMが消費すると推定されている。このためDRAMの消費電力削減は社会的にも非常に重要な課題である。

一方、コンピュータメモリシステムでは、1990年代終わりに、2次キャッシュがCPUチップに内蔵され始め、同時にパーフォーマンスを上げるために、DRAMベースの3次キャッシュを入れた階層構造を持つサーバーの研究も始まり、HP、IBMからこのような製品を使ったシステムが発表された。2次キャッシュまででキャッシュミスの確率を十分に減らすことができるので、3次キャッシュへの要件はアドレスアクセスが5～10nsで、データ転送速度が速いということである。

メモリ階層構造の下位レベルになるほど待機時間の割合が大きくなる。DRAMベースの3次キャッシュとメインメモリに高速でかつ待機時の消費電力が低減できる不揮発性メモリが使えるようになれば、メモリ階層構造やバックアップシステムが簡素化でき、メモリシステムのエネルギー消費量を削減できる。このため、3次キャッシュの候補として、高速の不揮発性メモリであるMRAM(magnetic random access memory)を取り上げ、可能性について検討した。

3次キャッシュとしてDRAM置き換えを考える場合、無限大の書き換え耐性があることの他に(1)DRAM並みに大容量化が可能なこと、(2)レイテンシ(latency)が小さく10ns以下のランダムアクセ

スが可能であること (3)動作時の消費電力が DRAM 並みに低いことが必要であり、種々のタイプのMRAMについて、実デバイスの試作及びシミュレーションを使ってこれらの観点から検討した。

1MbitMRAM は 0.18umAl2 層ロジックプロセスをベースに、書き込み用ワード線、MTJ(magnetic tunnel junction)、ビット線を追加して試作した。ビット線とワード線は Cu 配線を用いた。メモリセルサイズは MTJ の形状を変えて検討できるように $2.07um^2$ とこのデザインルールでは大きめにレイアウトした。メモリの構成は 16 I/O x 8k ワード x 8MAT で 1MAT 内にはワード線方向に 256 ビット、ビット線方向に 512 ビット並ぶ。セルサイズを縮小するために、MTJ と下部電極からの引き出し配線の自己整合構造、ボーダーレスビットコンタクト構造を考案した。セルサイズは従来のレイアウトのまま詰めると $27F^2$ (F はゲート配線の最小ピッチの 1/2 でデザインルールに依存しない。DRAM のセルサイズは $6F^2$～$7F^2$) になる。これは、汎用 DRAM の 4 倍、混載 DRAM の 1.5 倍のセルサイズである。また、MTJ から下地トランジスタの拡散層への配線を書き込み用ワード線と同層のランディングパッドを介さないで接続するレイアウト、書き込みワード線を貫通させるレイアウトを考案し、これで $13.5F^2$ まで縮小できることが分かった。

MRAM は単一のエネルギー障壁を持つ唯一のメモリであり、熱安定性(熱的に磁化反転が起こる)を考慮した設計が必要になる。アステロイド特性を使った書き込みを行う従来型の MRAM では、エネルギー障壁が最も低くなるのは半選択状態で、この状態を考慮したデバイス設計が必要である。800Mbps の高速データ転送を仮定して、メモリ LSI としての動作を考えると動作領域が非常に狭くなり、大容量の MRAM を実現することは難しいことを明らかにした。これに対して、トグル型書き込みは、動作点が広く、2ns の短パルスで書き込みできることを明らかにした。課題は低消費電力化であるが、外部からバイアス磁界を加えることにより、アステロイド書き込みと同レベルの 5mA 以下の電流で書き込みできることが分かった。

スピン注入磁化反転型メモリは、従来型の MRAM より書き込み電流を 1/10 以下に低減できる。短パルス書き込みとの組み合わせにより、DRAM と同等のエネルギーで書き込みができる。しかし、メモリセルの選択トランジスタが電流駆動トランジスタを兼ねており、トランジスタのレイアウトでセルサイズが決まるので、セルサイズ縮小には更なる書き込み電流低減(トランジスタのゲート幅縮小)が課題である。この電流低減と繰り返し書き込み信頼性に問題が無ければ、DRAM 置き換えの最も有力な候補になることを明らかにした。

MRAM の高速動作について、書き込みは数 ns 電流パルスでスイッチングでき、この動作は書き込み/読み出しサイクルの中に隠すことができるので問題にならない。読み出しについてはバイアス印加時の MR(磁気抵抗変化率：magnetoresistance) 比を 80%以上にするとビット線のプリチャージからセンスできるまでの時間が 1ns 以下になることを明らかにした。この結果から、5ns 以下での動作は実現できる見通しである。また、80%以上の MR 比は MTJ のトンネル酸化膜を結晶構造の MgO にすることによって、実現できる。

コンピュータシステムの低消費電力化では、動作時の消費電力削減も重要である。今まではコンピュータシステムは CPU、キャッシュ、メインメモリというように別チップで構成され PCB(printed circuit board)上に平面的に配置され接続されてきた。しかし、各チップの高速化が進むにつれ PCB 上の長配線による配線容量や配線抵抗、インダクタンス、及び各 LSI チップの I/O(input/output)部の容量による信号伝播遅延や寄生容量の充放電による動作時の消費電力増大が問題になってきた。これに対して、3 次元 LSI 技術を用いて CPU チップとキャッシュを積層にできれば、バンド幅を広く取れ、寄生容量も小さくなるため、動作時の消費電力も格段に低減できる。

本研究ではCPUチップとキャッシュを積層して、チップ面積の増加無く回路ブロック間を接続できる5μm以下のピッチのマイクロバンプ形成プロセスを検討した。新たな平坦化リフトオフプロセスを検討し、プロセス温度を制御して5μmピッチで2μmx2μmのマイクロバンプの試作に成功した。また、机上検討ではあるが、本技術によって4μmピッチまで縮小できる見通しであり、更なる微細ピッチ実現のための方向付けも行った。

論文審査結果の要旨

プロセッサとメモリ間で頻繁にデータのやり取りが行われる大規模なコンピュータシステムでは，システム性能を上げるためにキャッシュメモリの高速化，大容量化が重要な鍵となっている。しかし，１種類のメモリだけで，高速化と大容量化を両立することが難しいため，現在のキャッシュメモリは種類の違った SRAM(Static Random Access Memory)や DRAM(Dynamic Random Access Memory)を，１次キャッシュ，２次キャッシュ，３次キャッシュのように階層的に配置して，高速化と大容量化を見かけ上実現している。このような階層的メモリ構成では，下位の階層に行くほどメモリのアクセス速度が遅くなり，システム全体としての性能が制約されることとなる。本論文は，このようなキャッシュメモリのもつ課題を克服するために，高速キャッシュ SRAM を実現するための設計手法と，高速不揮発性メモリである MRAM(Magnetic Random Access Memory)の３次キャッシュとしての可能性について論じるとともに，アクセス速度を落とさずにメモリの階層性を高めるための３次元積層型キャッシュメモリモジュール用マイクロバンププロセスについて検討したもので，全編５章よりなる。

第１章は序論であり，本研究の背景と目的を述べている。

第２章では，高速キャッシュ SRAM を実現するためのデバイス，プロセス設計方法について論じるとともに，安定に動作する高速キャッシュ SRAM を実現するためには，半導体素子の寸法や特性のばらつきを考慮した設計，信頼性を考慮した設計が重要であることを述べている。特に，信頼性に関しては，SRAM のソフトエラーの主要因が熱中性子であることを突き止めて，ECC(Error Correction Code)回路により対処できることを明らかにしており，これらは重要な成果である。

第３章では，３次キャッシュとして用いられている DRAM を MRAM で置き換えることを提案している。MRAM を実際に試作，評価し，MTJ(Magnetic Tunnel Junction)のアステロイド特性を用いた書き込み方式では，半選択状態にあるメモリセルの動作領域が充分確保できないために，大容量化が難しいことを明らかにしている。動作領域を広く取れるトグル型書き込み方式の MRAM についても評価し，トグル型書き込み方式は低消費電力化に課題があることを明らかにしている。これらの評価結果を基に，DRAM 並みのメモリセル・サイズと SRAM 並みの高速性を有するスピン注入磁化反転型 MRAM が DRAM を置き換える可能性が最も高いと結論している。これらの結果は重要な知見である。

第４章では，３次元積層型キャッシュメモリモジュールを作製するためのマイクロバンププロセスを提案している。平坦化リフトオフ法という新しい手法を採用することにより，狭ピッチ(５μm)の In/Au 微細マイクロバンプ(２μm×２μm) から成るマイクロバンプアレイの形成に成功している。これらは重要な成果である。

第５章は結論である。

以上，要するに本論文は，従来のキャッシュメモリのもつ課題を克服するために，高速キャッシュ SRAM を実現するための設計手法と，３次キャッシュとしての MRAM の可能性を明らかにするとともに，メモリの階層性を高めるための３次元積層型キャッシュメモリモジュール用マイクロバンププロセスの有効性を明らかにしたもので，半導体工学およびバイオロボティクスの発展に寄与するところが少なくない。

よって，本論文は博士(工学)の学位論文として合格と認める。